International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Survey Paper on Real-Time Document Commendations Based on User Discussions

Pramila S. Gaidhani[1], Manisha Naoghare[2]

[1]Student, Master of Engineering, Department of Computer Engineering, Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

[2]Professor, Assistant Professor, Department of Computer Engineering, Sir Visvesvaraya Institute of Technology, Chincholi, Sinner

Abstract: ***This paper discourses the problem of keyword extraction from talks, with the goal of utilizing these watchwords to recover for each short conversation part, a little number of conceivably relatable reports, which can be prescribed to members, just-in-time. The purpose of meetings is to facilitate direct communication between participants. Document plays an important role in meetings. Documents contain facts that are currently discussed, but they are not necessarily at hand, the method known as keyword extraction and grouping is overviewed which impulsively recommend the documents that are related to user's current activities for an ongoing discussion. In any case, even a short audio fragment contains a mixed bag of words, which are conceivably identified with a few topics; also, utilizing Automatic Speech Recognition (ASR) framework slips errors in the output. Along these lines, it is hard to assumption correctly the data needs of the discussion members. Firstly propose a calculation to remove decisive words from the yield of an ASR framework (or a manual transcript for testing) to coordinate the potentially differing qualities of subjects and decrease ASR commotion. At that point, make use of a technique that to make many implicit queries from the selected keywords which will in return produce list of relevant documents. The scores demonstrate that our proposition moves forward over past systems that consider just word recurrence or theme closeness, and speaks to a promising answer for a report recommender.***

Keywords: Automatic Speech Recognition, Keyword Extraction, Document recommendation, Conditional Random Fields

1. Introduction

We all are surrounded by wealth of information which is available in the form of databases, documents, or multimedia resources. But even this availability, access to this is conditioned by the availability of search engines. Users do not start searching for the information because their current activity does not allow them to do the search or they are not aware that the related information is available. To solve this problem just in time retrieval system is adopted, which spontaneously recommend documents that are related to the current users activities. When these activities are conversational, for instance when users participate in a meeting, their information needs can be modeled as implicit queries which are constructed in the background from the pronounced words, obtained through real-time automatic speech recognition (ASR). These implicit queries are used to retrieve and recommend documents from the web or local repository [3]. The intent behind keyword extraction from conversations is to provide a set of words which are ideal to the semantic content of the conversation. Therefore the aim is to find set of keywords, clustering of keywords and present result of this query to users in the form of documents. Mainly topic-based clustering is used to reduce the scope of inclusion of ASR errors into the queries. The focus of this is on formulating implicit queries to a just-in-time retrieval system for use in meeting rooms. It is important that the keyword set retains the diversity of topics from the conversation. While the early keyword extraction methods ignored topicality of conversation as they were based on word frequencies, more recent methods have considered topic modeling factors for keyword extraction, but without specifically setting a topic diversity constraint, which is important for naturally occurring conversations [1]. Consider scenario of meeting where documents related to meeting discussion are already informed to participants of meeting. Due to some of the reasons participants does not have such client time to search that contents on the internet or on any other source of information. During meeting to find information related to some point is very difficult without interrupting the discussion. This problem occurs most of the time in meeting. To fulfill the information needs of participants some systems must be developed which will take conversation as query and give related documents to that without the direct interaction of participants to the system.

2. Related Works

Just-in-time retrieval systems have the potential to bring a radical change in the process of query based information retrieval. Such systems continuously monitor users activities to detect information needs, and pro-actively retrieve relevant information. To achieve this, the systems generally extract implicit queries (not shown to users) from the words that are written or spoken by users during their activities. In this section, review existing just-in-time-retrieval systems and methods used by them for query formulation. Also discuss previous keyword extraction techniques from a transcript. Human inter-actions with everyday productivity applications (e.g. word processors, Web browsers, etc.) provide rich contextual information that can be leveraged to support just-in-time access to task-relevant information. As evidence for claim, here present Watson a system which gathers contextual information in the form of the text of the document the user is manipulating in order to proactively retrieve documents from distributed information repositories. This system close by describing the results of several experiments with Watson [11], which shows it consistently, provides useful information to its users.

• **M. Habibi and A. Popescu-Belis, Enforcing topic diversity in a document recommender for conversations (2014):** In this system, recommend set of rules for diverse merging of those lists, using a sub modular praise characteristic that rewards the topical similarity of files to the verbal exchange phrases in addition to their diversity. Comparing the proposed method through crowd sourcing. The consequences show the prevalence of the various merging technique over numerous others which no longer enforce the range of topics.

Drawback: when juxtaposed in an implicit query, these topics may have noisy effects on the retrieval results.

• **Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta, The AMIDA automatic content linking device Just-in-time document retrieval in meetings (2008):** It is a just- in-time document retrieval system that constantly retrieves items from a repository and displays them to a participant or to all of them. The repository includes meeting related documents together with selection from previous meetings of the group. The device can be used online during a meeting, but also offline, integrated in a meeting browser.

A Speech-based Just-in-Time Retrieval System monitors an ongoing conversation or a monologue and enriches it with potentially related documents, including multimedia ones, from local repositories or from the Internet. The documents are found using keyword-based search or using a semantic similarity measure [9] between documents and the words obtained from automatic speech recognition. Results are displayed in real time to meeting participants, or to users watching a recorded lecture or conversation. Several methods have been proposed to automatically extract keywords from a text, and are applicable also to transcribed conversations. The earliest techniques have used a new keyword extraction algorithm that applies to a single document [8] without using a corpus. Frequent terms are extracted first, and then a set of co-occurrence between each term and the frequent terms, i.e. occurrences in the same sentences, is generated. Co-occurrence distribution shows importance of a term in the document as follows. If probability distribution of co-occurrence between the term and the frequent terms is biased to a particular subset of frequent terms, then the term is likely to be a keyword. Manual assignment high quality keywords is expensive, time-consuming, and error prone. Therefore, most algorithms and systems aimed to help people perform automatic keywords extraction have been proposed.

• **M. Habibi and A. Popescu-Belis, Diverse keyword extraction from conversations(2013):** An improved method for keyword extraction from conversations rewards both word similarity-to extract the most representative words, and word diversity-to cover several topics if necessary were introduced. Inspired from summarization, the method maximizes the coverage of topics; those are recognized automatically in transcripts of conversation fragments. But it was unable to retrieve documents. The present system addresses the problem of building concise, diverse and relevant lists of documents [2], which can be recommended to the participants of a conversation to fulfill their information needs without distracting them. These lists are retrieved periodically by submitting multiple implicit queries derived from the pronounced words. Each query is related to one of the topics identified in the conversation fragment preceding the recommendation, and is submitted to a search engine over the English Wikipedia. Here developed an algorithm for diverse merging of these lists, using a sub modular reward function that rewards the topical similarity of documents to the conversation words as well a their diversity.

The benefit of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. The future method for diverse keyword extraction proceeds in three steps,

1) Used to represent the division of the abstract subject for each word.
2) These topic models are used to determine weights for the abstract topics in each conversation fragment represented by
3) The keyword list W = {w1, w2……wk}. Which covers a maximum number of the most important topics is preferred by rewarding range, using a unique algorithm introduced in this part.

• **Automatic Keyword Extraction from Document Using Conditional Random Field:** Conditional Random Fields (CRF) model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labeling. Many unsupervised approaches. The AMI (DA) system is a meeting room speech recognition system that has been developed and evaluated in the context of the NIST Rich Text (RT) evaluations. Recently, the Distant Access requirements of the AMIDA project have necessitated that the system operate in real-time. Another more difficult requirement is that the system fit into a live meeting transcription scenario. The AMI system for meeting room recognition is a combination of beam-forming, divarication and ASR in real time .Document summarization algorithms are most commonly evaluated according to the intrinsic quality of the summaries they produce. An alternate approach is to examine the extrinsic utility of a summary, measured by the ability of the summary to aid a human in the completion of a specific task. This uses topic identification as a proxy for relevancy determination in the context of an information retrieval task, and a summary is deemed effective if it enables a user to determine the topical content of a retrieved document.

• **An Overview of the Technique Used for Extracting Keywords from Documents:** In this paper, they focus on one speech category the combined meeting domain. Meeting speech is significantly different from written text and most other speech data. For example, there are typically multiple participants in a meeting, the discussion is not well organized, and the speech is spontaneous and contains disfluencies and ill-formed sentences. It is thus questionable whether to adopt approaches that have been shown before to perform well in written text for automatic keyword extraction in meeting transcripts. This paper evaluates several different

keyword extraction algorithms using the transcripts of the ICSI meeting corpus.

3. Exiting Techniques of Keyword Extraction

Various methods of locating and determining keywords have been used, both individually and in concert. In spite of their differences, a large amount of methods have the equal purpose and try to do the same thing: using some heuristic (such as distance between words, frequency of word use, or predetermined Word relationships), locate and define a set of words that accurately convey themes or describe information contained in the text.

- **Word Frequency Analysis**

A lot before time work troubled the frequency of term usage in the content, except the majority of this work focused on defining keywords in relation to a single document during 1972, the thought of statistically analyzing the frequency of keyword usage surrounded by a document in relative to multiple other documents became more common. This method, recognized as Term Frequency Inverse Document Frequency or purely TF-IDF, weights known term to conclude how well the term describes an individual paper inside a corpus. It does this by weighting the term positively for the number of times the term occurs within the specified document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term t and document d 2 D, where t appears in n of N documents in D. The TF-IDF as follows:

$$TFIDF(t, d, n,N) = TF(t, d)\ IDF(n,N)$$

- **Word Co-Occurrence Relationships**

While many methods of keyword extraction rely on word frequency (either within the document, within the corpus, or some combination of these), various possible problems have been pointed out with these metrics ,Including reliance on a corpus, and the assumption that a good keyword will appear frequently within the document but not within other documents within the corpus. These techniques to do not try to monitor any kind of relationship among words in a document.

4. Conclusion

As this complete paper describe different approaches on keyword extraction, but none of the approaches are seems to be perfect. This paper discourses the problem of keyword extraction from talks, with the goal of utilizing these watchwords to recover for each short conversation part, a little number of conceivably relatable reports, which can be prescribed to members, just-in-time.

5. Acknowledgement

I would like to express my profound gratitude and deep regard to my Project Guide Prof. M.M. Naoghare, for her exemplary counsel, valuable feedback and constant fillip throughout the duration of the project. Her suggestions were of immense help throughout my project work. Working under her was an extremely knowledgeable experience for me.

References

[1] M. Habibi and A. Popescu-Belis, “Keyword Extraction and Clustering for Document Recommendation in Conversations” IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,VOL. 23,NO. 4, APRIL 2015.

[2] M. Habibi and A. Popescu-Belis, “Enforcing topic diversity in a document recommender for conversations,” in Proc. 25th Int. Conf.Comput.Linguist.(Coling), 2014, p588–599.

[3] M. Habibi and A. Popescu-Belis, “Diverse keyword extraction from conversations,” in Proc. 51st Annu.Meeting Assoc. Comput. Linguist.,2013, pp. 651–657.

[4] D. Harwath and T. J. Hazen, “Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,”in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012,pp. 5073–5076.

[5] F. Liu, D. Pennell, F. Liu, and Y. Liu, “Unsupervised approaches for automatic keyword extraction using meeting transcripts,” in Proc. Annu. Conf. North Amer. Chap.ACL(HLT-NAACL),2009,pp.620–628.

[6] P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiát, D.Korchagin, M. Lincoln, V. Wan, and L. Zhang, “Real-time ASR from meetings,”in Proc. Interspeech, 2009, pp. 2119–2122.

[7] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and Wang,“Automatic keyword extraction from documents using conditional random fields,”J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.

[8] Y. Matsuo and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information,” Int. J. Artif.Intell.Tools, vol. 13, no. 1, pp. 157–169, 2004.

[9] A. Popescu-Belis, M. Yazdani, A. Nanchen, and P. N.Garner, “A speech-based just-in-time retrieval system using semantic search,” in Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL), 2011, pp. 80–85.

[10] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S.Castronovo, T.Wilson, A. Jaimes, and J. Carletta, “The AMIDA automatic content linking device: Just-in-time document retrieval in meetings,” in Proc.5th Workshop Mach. Learn. Multimodal Interact. (MLMI), 2008, pp.272–283.

[11] J. Budzik and K. J. Hammond, “User interactions with everyday applications as context for just-in-time information access,” in Proc. 5th Int. Conf. Intell. User Interfaces (IUI’00), 2000, pp. 44–51.

[12] P. E. Hart and J. Graham, “Query-free information retrieval,” Int.J. Intell. Syst. Technol. Applicat., vol. 12, no. 5, pp. 32–37, 1997

[13] B. J. Rhodes and P. Maes, “Just-in-time information retrieval agents,”IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

[14] M. Habibi and A. Popescu-Belis, “Using crowdsourcing to compare document recommendation strategies for

conversations," Workshop Recommendat. Utility Eval.: Beyond RMSE (RUE'11), pp. 15–20,2012.
[15] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. New York, NY, USA: Springer, 2012, ch. 3, pp. 43–76.
[16] J. Wang, J. Liu, and C. Wang, "Keyword extraction based onpagerank," in *Proc. Adv. Knowl. Disc. Data Mining (PAKDD)*, 2007,pp. 857–864
[17] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proc. Conf. Empir. Meth. Nat.Lang. Process. (EMNLP'10)*, 2010, pp. 366–376.
[18] K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A keyphrase based approach to interactive meeting summarization," in Proc. IEEE Spoken Lang. Technol. Workshop (SLT'08), 2008, pp. 153–156.
[19] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL), 2009, pp. 620–628
[20] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields," J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.

Author Profile

Pramila Gaidhani received the B.E. degree in Computer Engineering from Maharashtra Institute of Technology in 2011. She is currently pursuing Master's Degree in Computer.

M. M. Naoghare is working as Assistant Professor in Sir Visvesvaraya Institute of Technology.